

知识管理论坛
2017 年 第 1 期（总第 7 期）

【学术探索】

基于关联数据的机构知识库联盟构建研究

◎ 陈和

厦门大学图书馆 厦门 361005

摘要：[目的 / 意义] 探讨采用语义网的方法实现区别于传统的构建机构知识库联盟的方法。[方法 / 过程] 把现有的机构知识库内容发布为关联数据，并在此基础上采用语义网的方法构建基于关联数据的机构知识库联盟。[结果 / 结论] 基于关联数据的机构知识库联盟可以实现资源浏览、检索等基本的功能，同时也可以采用语义扩展、语义推理等方法实现传统联盟无法实现的知识服务功能。构建基于关联数据的机构知识库联盟是语义网的一个具体实践，可以提供比传统联盟更多更好的功能服务，而且构建的成本投入少，见效快。

关键词：机构知识库　关联数据　资源联盟　语义网　知识服务

分类号：G203

引用格式：陈和 . 基于关联数据的机构知识库联盟构建研究 [J/OL]. 知识管理论坛 , 2017, 2(1): 2-8[引用日期]. http://www.kmf.ac.cn/p/1/87/.

1 机构知识库联盟发展现状

机构知识库（institutional repository，IR）经过最近几年的发展，在国内外都获得了长足的进步。从数量上来看，根据 OpenDOAR（the Directory of Open Access Repositories，开放获取知识库目录）站点统计，截至 2016 年 6 月 1 日，注册登记的机构知识库数量达到 3 100 个 [1]；在 CALIS（China Academic Library & Information System，中国高等教育文献保障系统）机构知识库站点上登记的国内大陆高校机构知识库有 40 个 [2]。实际上，机构知识库数量远不止这些，因为还有很多机构知识库没有在这些网站上登记。从 OpenDOAR 站点的机构库数量统计图上可以看出，随着时间的推移，机构知识库数量仍会保持一定的速度持续增长。

随着机构知识库数量的增多，以及资源内容的不断增长，实现机构之间资源共建共享的需求越来越强烈。机构知识库联盟作为资源共建共享的一种方式也应运而生，有学者认为机构知识联盟是机构知识库的未来发展趋势之一。通过构建机构知识库联盟，可以展现联盟整体的学术研究成果；可以在更大范围内进行知识产出保存和共享；可节约成本，体现规模效益；可促进单个机构知识库的建设推广；是构建全国知识基础保存设施的重要步骤 [3-4]。

截至目前，国内外已经构建了不少的机构知识库联盟，既有国家层面的联盟，也有区域性的联盟，比如有澳大利亚的 ARROW[5]、日本的 JAIRO[6]、欧盟的 DRIVER[7]、英国的白玫瑰知识库联盟 [8]、美国的 ALADIN 联盟 [9]、俄

作者简介：陈和（ORCID：0000-0002-4506-5612），副研究馆员，硕士，E-mail：xmu_chen@xmu.edu.cn。

收稿日期：2016-10-18　　发表日期：2017-01-18　　本文责任编辑：王传清

知识管理论坛，2017（1）:2–8
DOI：10.13266/j.issn.2095–5472.2017.002

亥俄州数字知识库联盟[10]、台湾学术机构典藏[11]、香港地区 HKIR[12]、中国科学院文献情报中心机构知识库[13]，以及 CALIS 机构知识库[2]等。

纵观目前的机构知识库联盟，其实现资源联盟的技术方式主要包括如下 3 种[14-15]：

（1）集中方式。由联盟中心统一构建、运行及维护知识库系统，联盟成员将数字资源上传到此统一系统中，并在此系统上为用户提供浏览、检索和下载等服务。联盟成员不必再承担构建及维护系统的任务，节约了成本。此联盟方式的代表有白玫瑰知识库联盟、ALADIN 联盟等。

（2）分布方式。联盟成员根据共同的数据交互标准和协议分别构建和维护各自独立的机构知识库。联盟中心按照数据交互标准和协议采集成员的元数据到中心平台，但原始数据仍然保留在成员机构知识库之中。在联盟中心平台上为用户提供统一浏览和检索服务，实现一站式服务。这是比较常见和主流的联盟方式，ARROW、JAIRO、俄亥俄州数字知识库联盟、台湾学术机构典藏、香港地区 HKIR、中国科学院文献情报中心机构知识库、CALIS 机构知识库等均为分布式联盟。

（3）综合方式。联盟方式既有集中方式，又有分布方式，其中机构知识库联盟的分布构建方式中的数据交互标准和协议基本上是采用 Dublin Core 元数据作为描述资源的元数据标准，采用 OAI-PMH 方式进行资源元数据收割。其工作方式是在联盟中心设置资源收割服务器，主动请求收割联盟成员机构知识库的元数据，然后对获取的元数据进行集中存储，在此基础上，构建联盟统一服务平台，为联盟成员提供资源统一浏览、检索、订阅、下载等信息服务。另一方面，这种通过数据收割方式构建的联盟，需要专门配置资源收割服务器、索引服务器、前端服务平台服务器，以及元数据资料存储空间等。在资源收割的过程中，常常受到联盟成员数量、资源内容数量、资源内容数据质量，以及通讯网络方面的影响。DRIVER 联盟就是采用的此方式。

各个机构单独构建的机构知识库通过上述方式组建成资源联盟后，摆脱了“信息孤岛”、服务资源有限的窘境，实现了资源统一揭示和利用，并在一定程度上实现了联盟成员之间资源的共建共享。然而，这种机构知识库联盟，其资源只是限于成员机构知识库中的资源，与知识库外界资源是绝缘的，不能与外界资源进行关联和交互，而且，联盟资源的服务对象是人，忽略了机器。这与机构知识库的真正开放精神还存在一些差距。

2 关联数据与机构知识库

2.1 关联数据概述

关联数据（linked data）最早由 T. Berners-Lee 于 2006 年 7 月首次提出，目前成为被 W3C（world wide web consortium，国际互联网协会）推荐的一种用来发布和联接各类数据、信息和知识的规范。关联数据采用 RDF（resource description framework，资源描述框架）数据模型，利用 URI（uniform resource identifier，统一资源标识符）命名数据实体，发布和部署实例数据和类数据，从而可以通过 HTTP 协议提示并获取这些数据，同时它强调数据的相互关联、相互联系和有益于人机理解的语境信息。

对关联数据进行访问或查询，W3C 推荐采用 SPARQL 查询语言，它是一种从 RDF 图获取信息的查询语言。

关联数据的构建和实现需要遵行 4 个基本原则[16-17]：①使用 URI 作为任何事物的标识名称；②使用 HTTP URI 让任何人都可以访问这些标识名称；③当有人访问某个标识名称时，提供有用的信息；④尽可能提供相关的 URI，以使人们可以发现更多的事物。

关联数据对数据访问方式进行了标准化操作，用户或代理无需知道某具体关联数据发布网站的体系架构、存储方式等任何技术细节，只要知道 Web 服务器地址，都可以直接用

SPARQL 进行访问。作为一种数据发布技术，由于支持语义描述，同时提供标准的服务接口，其有效地提高了数据的可查找性和重用性。

2.2 机构知识库发布为关联数据的优势分析

机构知识库作为机构成员学术作品长期保存和统一展示的学术信息平台，采用了标准的资源组织方式和资源著录方式，具备转换为关联数据的先天优势，归纳如下：

（1）句柄系统。多数机构知识库构建软件都使用句柄系统来标识内容对象，比如 DSpace 采用 CNRI 的 handle 系统，每一个对象被赋予全球唯一的可供识别和访问的 handle 标识 [18-19]。也有机构知识库是采用 DOI 作为句柄系统 [20]。

（2）开放访问。机构知识库的资源内容一般都遵循开放获取政策，在 Web 上，用户可以通过浏览器和全球唯一的句柄号访问对应对象的详细信息。

（3）统一元数据标准。为了方便系统之间的数据交互，多数机构知识库构建软件采用了统一的元数据标准，而且是以 Dublin Core 元数据标准为主，而 Dublin Core 在关联数据中是一个重要的领域本体。

（4）资源组织方式。机构知识库通常是按照“机构－部门－条目”来组织资源内容，这种方式默认潜在地包含了资源的归属和层级约束关系。

因此，比照关联数据的四原则，现有的机构知识库已经具有了关联数据的部分特征，具备把现有机构知识库内容发布为关联数据的条件。

3 基于关联数据的机构知识联盟实现方式

可以通过两步来实现基于关联数据的机构知识库联盟，即首先把各个机构知识库内容发布为关联数据，然后在此基础上构建联盟服务系统。

3.1 发布关联数据

现有的机构知识库构建软件底层基本都采用关系型数据库，因此，可以通过关联数据技术直接把关系型数据库数据发布为关联数据。目前可以实现的技术有 VirtucsoUniversal Server、Triplify、D2RQ 等，其中 D2RQ 技术是推荐的主流方案 [21]。如果机构知识库构建软件是 DSpace，只需要进行简单的安装与配置，即可把对象数据发布为关联数据。下面分别对这两种实现方式进行介绍。

3.1.1 D2RQ 发布关联数据

D2RQ 能够将关系数据库抽象为只读的 RDF 图，而且无需将数据拷贝为 RDF 存储，直接提供基于 RDF 的数据访问的开源软件平台。D2RQ 平台由如下 3 个组件构成 [22]：

（1）D2RQ 映射语言。定义将关系型数据转换成 RDF 格式的映射规则。

（2）D2RQ Engine。使用 D2RQ 映射文件将关系型数据库中的数据映射成虚拟的 RDF 数据（实际上并没有将关系型数据库发布成真实的 RDF 数据）。当访问关系型数据时将 RDF 数据的查询语言 SPARQL 转换为关系型数据的查询语言 SQL，并将 SQL 查询结果转换为 RDF 三元组或者 SPARQL 查询结果。

（3）D2R Server。它是一个 HTTP Server，主要提供对 RDF 数据的查询访问接口，以供上层的 RDF 浏览器、SPARQL 查询客户端以及传统的 HTML 浏览器调用。三者关系如图 1 所示：

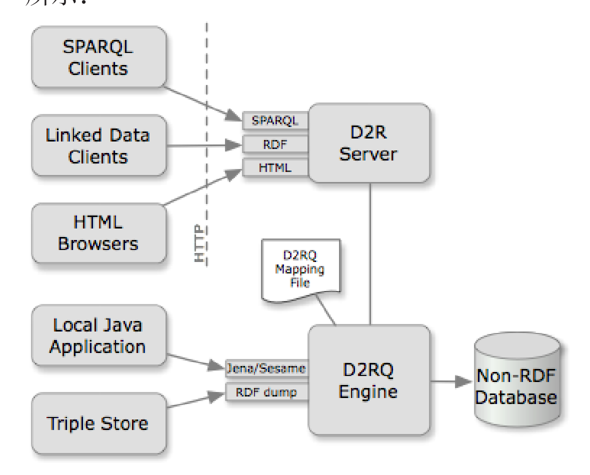


图 1 D2RQ 结构功能示意 [23]

知识管理论坛，2017（1）:2–8
DOI：10.13266/j.issn.2095–5472.2017.002

D2RQ 发布关联数据方法步骤如下[24]：

（1）从官网上下载并解压 D2RQ 软件到合适目录。

（2）下载对应关系型数据库的驱动程序 JAR 包，置于 [d2rq-source]/lib/db-drivers/ 目录下，并注意引用该程序的完整类名（如：org.postgresql.Driver）和 JDBC URL 模式（如：jdbc:postgresql://servername:5432/database）。

（3）生成 D2RQ 映射文件。

（4）启动 D2RQ 服务。

上述第（3）步骤最为关键和重要，需在命令模式下，切换到 [d2rq-source] 目录，使用命令 generate-mapping 生成映射文件。使用 generate-mapping 命令时，必须指定连接数据库的用户名和密码，需要发布为关联数据的数据库名、数据表名、字段名等，以及需要忽略或跳过的数据表名、字段名等。

D2RQ 服务启动之后，可以在本机浏览器上输入如下地址进行访问和测试：http://localhost:2020，此时，对外的 SPARQL 端点地址为：http://localhost:2020/sparql。

3.1.2 DSpace 发布关联数据

自从 5.0 版本开始，DSpace 支持把仓储系统中的数据发布为关联数据，使其在支持 OAI-PMH 的基础上，又增加了一条分享内容的渠道。

DSpace 把对象内容发布为关联数据的方法步骤[25]：

（1）修改配置文件：

① [dspace]/config/dspace.cfg

在参数变量"event.dispatcher.default.consumers"之后添加"rdf"。

② [dspace]/config/modules/rdf.cfg

设置参数变量"public.sparql.endpoint"和"contextPath"，分别设置成需要的值。

（2）下载并开启 jena-fuseki 服务：在 linux 下，开启 jena-fuseki 服务的命令如下：

export FUSEKI_HOME=[path_to_fuseki]/jena-fuseki;$FUSEKI_HOME/fuseki-server --localhost --config[dspace]/config/modules/rdf/fuseki-assembler.ttl

执行上述命令后，即开启了 fuseki 服务，若是第一次执行，会在 [dspace] 目录下生成 triplestore 目录，与 RDF 相关的文件便存储在此目录下。

（3）生成 RDF 文件：开启 fuseki 服务后，可以把 DSpace 系统中的已有的对象数据转化为 RDF 数据（三元组形式），转化的操作命令如下：

[dspace]/bin/dspacerdfizer -c –v

执行如上命令后，DSpace 系统中的所有可读的条目将转化为 RDF 数据。之后新上传到系统中的条目数据将自动转化为 RDF 数据。

（4）部署 RDF 虚拟目录：在安装 DSpace 时若没有部署过 rdf，此时需要重新部署，需要拷贝 [dspace]/webapps/rdf 目录到 [tomcat]/webapps/ 目录下。

重新启动 tomcat 服务后，便可提供关联数据服务，在本机浏览器上可输入如下地址进行访问和测试：http://localhost:3030/，此时，对外的 SPARQL 端点地址为：http://localhost:3030/dspace/sparql。

因此，基于 DSpace 构建的机构知识库，只要把 DSpace 系统升级到 5.0 以上版本后，根据上述方法步骤进行操作，即可发布关联数据。

3.2 基于关联数据的联盟服务系统

联盟服务系统是统一为成员用户提供服务的平台，与传统的联盟服务相似，包括统一检索服务、资源浏览服务、成员注册登记等，但其实现方式却不同于传统的联盟。

3.2.1 统一检索服务

基于关联数据的联盟不需要收割各成员机构知识库的元数据，不需要集中存储元数据。其统一检索服务是基于各个成员机构知识库提供的标准 SPARQL 端点来实现。具体可以采用开源软件 DARQ 来实现。

DARQ 是一个基于 SPARQL 的联邦查询搜索引擎，是在 JenaARQ[26] 的基础上增加了一个新的查询设计算法，并修改查询执行引擎扩

展而来，为不同的、分布的 SPARQL 端点提供透明查询访问。对用户而言，使用 DARQ 查询多个 SPARQL 端点就好像只查询一个 SPARQL 端点，它只提供一个查询接口，而把对多个 SPARQL 端点的复杂查询的联邦细节都留给了查询搜索引擎后台处理 [27]。

DARQ 的具体使用类似 Jena ARQ 的用法，但相对比较简单。关键是需要预先设置好服务描述配置文件（service descriptions），定义好各个 SPARQL 端点的服务能力，比如 SPARQL 端点的 URL 地址、数据描述、访问模式、结果返回数量等等。

3.2.2 数据浏览

各成员机构知识库发布关联数据后，用户可以通过语义浏览器（semantic web browsers），如 Disco[28]、Tabulator[29]、OpenLinkData 扩展 [30] 等，浏览各成员的知识库中的 RDF 数据。由于关联数据的特性，用户可以跟随 RDF 语句表达的链接，在不同的数据源之间浏览，可以从一个数据源开始逐渐遍历整个网络。如果原来机构知识库已经进行关联语义扩展，将会浏览到更丰富的内容。这种浏览方式有别于传统的文档链接浏览，将带给用户不一样的浏览体验。

3.2.3 成员注册与登记

为了便于联盟成员之间或者成员之外发现和利用各个成员的机构知识库数据，需要把成员机构知识库的相关信息进行注册和登记。登记的信息需要遵循 W3C 推荐的 VoID 标准 [31]，它是描述 RDF 数据集的元数据，包括一般性元数据（general metadata）、访问元数据（access metadata）、结构元数据（structural metadata）和数据集间关联描述（description of links between datasets）。使用 VoID 标准登记信息是为了便于人们为了不同的任务而找到有用的数据。

3.2.4 知识服务

基于关联数据的机构知识库联盟其实是一个轻量级的语义网络，可以利用语义关联、语义挖掘和推理的方法发现有用的信息。例如可以把作者的论文和科研数据进行关联，为用户展示整个科研生命周期的数据生成、数据分析和科研成果等；可将某一知识主题的相关内容进行知识聚合，以期发现新的知识内容；可将相关内容数据之间的关联关系建立知识地图，并利用可视化的技术表现出来，实现知识图谱等。所以在联盟服务端需要提供多种知识发现模式，和 / 或者 SPARQL 查询语句可编辑器，方便用户设置不同的知识发现模式，结合机器学习，挖掘和发现机构知识库联盟中有用的知识。

4 分析与讨论

以上简要介绍了把现有机构知识库内容发布为关联数据，并在此基础上构建联盟服务的方法。从整个构建过程看，有其优势，也有其不足。

优势在于：①万维网逐渐扩展到语义网，是一种发展趋势，而基于关联数据的机构知识库联盟正是在这个趋势下的一个具体实践，可以探索和积累万维网向语义网转变的方法和经验。②基于关联数据的机构知识库联盟，是一种轻量级的语义网，可以借助语义网的理论与技术来充分利用机构知识库中的数据，发掘机构知识库数据的潜在价值，同时推进语义网理论创新和技术进步。③基于关联数据的机构知识库联盟，给用户带来不一样的使用体验和感受，将促进机构知识库的内容建设和系统转型升级。④联盟服务系统由于不需要收割元数据，也不需要集中存储元数据，所以不需要额外开发收割软件，不需要部署收割服务、存储服务器以及索引服务器等，节省了元数据收割和建立索引以及系统维护的时间，费效比比传统分布式联盟高出许多。

不足在于：①在构建过程中，有一个比较大的问题在数据质量方面。目前的机构知识库内容包括了期刊论文、会议论文、工作文稿、专利文献、学位论文、图书章节等，由于考虑到数据加工、数据交互等方面的问题，采

知识管理论坛，2017（1）:2–8
DOI：10.13266/j.issn.2095–5472.2017.002

用的元数据标准基本上都是 Dublin Core 元数据。众所周知，Dublin Core 元数据标准简洁轻便，容易使用，但是其优点也导致了其不足，表现为对资源的描述不够深入，比如没有专门的元数据描述期刊论文的参考文献部分，使得论文之间的相互参照关系缺失，会议论文没有会议相关的属性描述，学位论文没有学生与指导教师的关系，以及缺失学位相关的属性等。关联数据注重实体的属性及之间的关系，现有的机构知识库资源转化关联数据后，实体的属性和关系不够丰富，进行数据推理时，达不到预想的结果。因此机构知识库发布为关联数据时，需提高或加强数据质量，王思丽、杨雪梅等提出的思路和方法或许可以提供参考，即首先抽取机构知识库内的核心实体类，利用 RDF 进行语义标注和关联，关联到外部的 DBpediaOntology、DBLPBibliography 等其他关联数据集，丰富语义内容，然后再发布为关联数据 [32-33]。更进一步地，在发布关联数据前还可以关联到机构学者中心、项目成果中心等数据源，进一步扩展机构知识库的语义内容。②关联数据的相关应用技术，还有待完善和创新。比如前述构建基于关联数据的机构知识库联盟服务系统中用到的 DARQ 软件，在检索效率方面还乏善可陈，特别是在 SPARQL 端点比较多时，检索效率比较低下。再比如语义浏览器，有些多年未更新，其功能还是比较单一、粗糙，有待加强和完善，在使用体验上还待提高。造成这种局面的原因主要是因为关联数据还处于发展初期，宣传推广不足，参与开发与研究的人员不多，缺乏投入等。

5 结语

把机构知识库内容发布为关联数据，并在此基础上构建机构知识库联盟，实现成员之间资源共建共享的目的。联盟不但提供传统的资源检索、浏览等服务，而且还提供可定制的知识服务。基于关联数据的联盟是开放的，不但联盟成员之间可以互联，联盟也可以与其他关联数据集进行互联，实现了真正意义上的开放，为机构知识库的应用提供了更广阔前景。本文对此联盟方式只是在技术实现上进行了探讨，尚未进行正式的实践，希望在今后的具体实践中进一步探索。

参考文献：

[1] Open DOAR[EB/OL]. [2016-09-25]. http://www.opendoar.org/.
[2] CALIS 机构知识库 [EB/OL]. [2016-09-25]. http://ir.calis.edu.cn/.
[3] 曾苏 , 马建霞 , 祝忠明 . 机构知识库联盟发展现状及关键问题分析 [J]. 图书情报工作 , 2009, 53(24): 106-110.
[4] 陈和 . 机构知识库发展趋势探析 [J]. 图书情报工作 , 2012, 56(21): 62-66.
[5] ARROW[EB/OL]. [2016-09-25]. http://arrow.edu.au/.
[6] JAIRO[EB/OL]. [2016-09-25]. http://jairo.nii.ac.jp/en/.
[7] DRIVER[EB/OL]. [2016-09-25]. http://www.driver-repository.eu/.
[8] WRCER[EB/OL]. [2016-09-25]. http://eprints.whiterose.ac.uk/.
[9] ALADIN[EB/OL]. [2016-09-25]. http://aladinrc.wrlc.org/.
[10] OhioLINK digital resource commons[EB/OL]. [2016-09-25]. http://drc.ohiolink.edu/.
[11] 台湾学术机构典藏 [EB/OL]. [2016-09-25]. http://tair.org.tw/.
[12] Hong Kong institutional repositories[EB/OL]. [2016-09-25]. http://hkir.ust.hk/hkir/.
[13] 中国科学院文献情报中心机构知识库 [EB/OL]. [2016-09-25]. http://ir.las.ac.cn/.
[14] 陈雨杏 . 我国区域机构知识库联盟的构建模式选择与实施策略 [J]. 图书馆学研究 , 2011(7): 59-63, 75.
[15] 陈和 . 构建国家机构仓储联盟探析 [EB/OL]. [2016-09-25]. http://www.paper.edu.cn/html/releasepaper/2010/01/5/.
[16] Berners-LeeT. Linked data[EB/OL]. [2016-09-25]. http://www.w3.org/DesignIssues/LinkedData.html.
[17] 刘炜 . 关联数据 : 概念、技术及应用展望 [J]. 大学图书馆学报 , 2011(2): 5-12.
[18] Dspace[EB/OL]. [2016-09-25]. http://www.dspace.org/.
[19] Handle.Net registry[EB/OL]. [2016-09-25]. http://www.handle.net/.
[20] Digital object identifier system[EB/OL]. [2016-09-25]. https://www.doi.org/.
[21] 濮德敏 , 任瑞娟 , 米佳 , 等 . 关系型本体转换为关联数据技术方案比较研究 [J]. 图书馆理论与实践 , 2014(12): 30-34.
[22] 如何利用 D2R 发布 Linked data[EB/OL]. [2016-09-



25]. https://www.ibm.com/developerworks/cn/web/1003_zhangjing_d2r/.

[23] The D2RQ platform[EB/OL]. [2016-09-25]. http://d2rq.org/.

[24] Getting started with D2RQ[EB/OL]. [2016-09-25]. http://d2rq.org/getting-started.

[25] Linked (open) data -D Space 5.x documentation [EB/OL]. [2016-09-25]. https://wiki.duraspace.org/display/DSDOC5x/Linked+(Open)+Data.

[26] Apache- ARQ - A SPARQL processor for Jena[EB/OL]. [2016-09-25]. https://jena.apache.org/documentation/query/.

[27] DARQ -Federated queries with SPARQL[EB/OL]. [2016-09-25]. http://darq.sourceforge.net/.

[28] Disco – Hyperdata Browser[EB/OL]. [2016-09-25]. http://wifo5-03.informatik.uni-mannheim.de/bizer/ng4j/disco/.

[29] Tabulator: generic data browser[EB/OL]. [2016-09-25]. https://www.w3.org/2005/ajar/tab.

[30] OpenLink data explorer extension[EB/OL]. [2016-09-25]. http://ode.openlinksw.com/.

[31] Vocabulary of interlinked datasets[EB/OL]. [2016-09-25]. https://www.w3.org/TR/void/.

[32] 王思丽 , 祝忠明 . 利用关联数据实现机构知识库的语义扩展研究 [J]. 现代图书情报技术 , 2011(11): 17-23.

[33] 刘雪梅 . 基于关联数据的机构知识库服务模式构建与实现 [J]. 图书馆学刊 , 2016(4): 4-7.

Research on Building Confederation of Institutional Repository based on Linked Data

Chen He

Xiamen University Library, Xiamen 361005

Abstract: [Purpose/significance] This paper aims to explore the method of using semantic web to build confederation of institutional repository which is different from the traditional way. **[Method/process]** It publishes the institutional repository content as linked data firstly, and uses the method of the semantic web to build confederation of institutional repository. **[Result/conclusion]** The confederation of institutional repository based on linked data can realize the basic functions such as browsing and retrieving. At the same time, the confederation can realize knowledge service function by extending semantic content and semantic reasoning. Building confederation of institutional repository based on linked data is a specific practice of semantic web, the confederation can provide more and better functional services than the traditional confederation, and the construction cost is low.

Keywords: institutional repository linked data resources confederation Semantic Web knowledge services